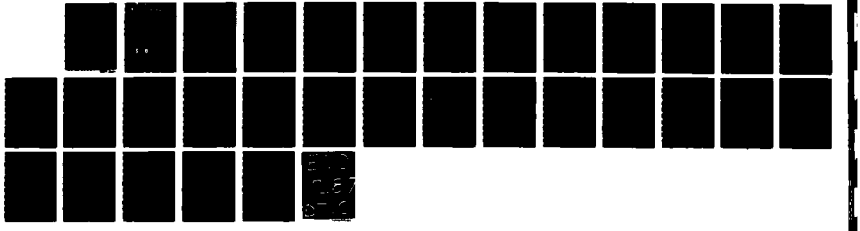
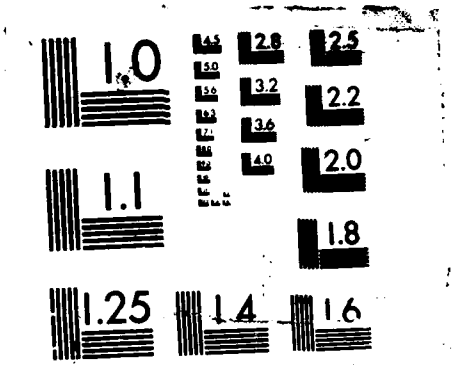
AD-A181 335 MOLECULAR BIOLOGY CONFERENCE ON GENETIC ENGINEERING 1/1
TECHNIQUES (2ND) HEL (U) OFFICE OF NAVAL RESEARCH
LONDON (ENGLAND) C F ZOMZELY-NEURATH 27 MAY 87
UNCLASSIFIED ONRL-7-809-C F/G 6/2 NL





(2)

ONRL Report 7-009-C

Molecular Biology: Conference on Genetic Engineering Techniques

Claire E. Zomzely-Neurath

27 May 1987

Approved for public release; distribution unlimited

U.S. Office of Naval Research, London

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS AD-A181 335
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) 7-009-C	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION US Office of Naval Research Branch Office, London 6b. OFFICE SYMBOL (If applicable) ONRL	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Box 39 FPO, NY 09510	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

Molecular Biology: Conference on Genetic Engineering Techniques

12. PERSONAL AUTHOR(S)

Claire F. Zomzely-Neurath

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Conference	FROM _____ TO _____	27 May 1987	26

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	
06	16		Molecular biology Genetics Bioengineering
06	03	02	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The topics covered at this conference included the synthesis of foreign products in *E. coli*, expression of cloned genes in yeast and cultured mammalian cells, the introduction of cloned genes into whole animals and plants, and studies on a number of specific genes which have a significant clinical potential. Presentations on these topics are summarized.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL C.J. Fox	22b. TELEPHONE (Include Area Code) (44-1) 409-4340 / 22c. OFFICE SYMBOL 11

CONTENTS

		Page
1	INTRODUCTION	1
2	ISOLATING LARGE SEGMENTS OF CHROMOSOMAL DNA	1
	Choice and Use of Cosmid Vectors	1
	Molecular Techniques in Mammalian Genetics: Techniques for Chromosome Walking	3
3	EXPLOITING *E. COLI*	7
	The Synthesis of Fusion Proteins in *E. coli* and Their Use in Raising Antibodies	7
	The Purification of Foreign Polypeptides Expressed in *E. coli*	9
	The Protein Engineering of Subtilisin	13
4	EXPRESSION IN MAMMALIAN CELL CULTURES AND TRANSGENIC MICE	14
	Eukaryotic Expression Vectors	14
	The Production of Transgenic Mice by the Direct Microinjection of Cloned DNA's	17
5	EXPRESSION AND SECRETION OF FOREIGN POLYPEPTIDES IN YEAST	21
6	ENGINEERING SPECIFIC GENES	23
7	CONCLUSON	25
8	REFERENCES	25

MOLECULAR BIOLOGY: CONFERENCE ON GENETIC ENGINEERING TECHNIQUES

1 INTRODUCTION

The Second European seminar on genetic engineering techniques, sponsored by IBC Technical Services Ltd., was held at the Portman Hotel, London, UK, 20 to 21 November 1986. The 60 participants from seven different European countries as well as the UK and US represented both academic and industrial organizations in a 1:1 ratio.

This interesting and intensive conference on genetic engineering techniques dealt with the continuing developments in these techniques which have led to remarkable progress in our understanding of gene organization and expression. This technology has had a major impact not only in academic research but also in industry where the potential of having a microorganism synthesize gene products of industrial and pharmacological importance was quickly realized. The topics covered at this conference included the synthesis of foreign products in *E. coli*, expression of cloned genes in yeast and cultured mammalian cells, the introduction of cloned genes into whole animals and plants, and studies on a number of specific genes which have a significant clinical potential.

2 ISOLATING LARGE SEGMENTS OF CHROMOSOMAL DNA

Choice and Use of Cosmid Vectors

Cosmid vectors are potentially the most useful DNA cloning vehicles because of their large capacity (40,000 base pairs [bp]), but their utility has been severely reduced because of problems of instability and irreproducibility of growth. P. Little (Institute of Cancer Research, Chester Beatty Laboratories, London, UK) discussed new vectors which may overcome many of these technical problems.

The application of recombinant DNA (rDNA) techniques to the study of human genetic disease has had a revolutionary effect upon our understanding of levels of gene dysfunction. When the affected gene can be identified, detailed analysis of DNA sequence has been possible, identifying specific mutations and allowing a clearer understanding of the nature of the diseased state. However, a major effort of human genetics is now devoted to the isolation of genes that are of unknown function but that must be responsible for some very common and serious genetic diseases. For example, Huntington's Chorea, Cystic Fibrosis, and Duchenne Muscular Dystrophy, amongst others, represent this class of gene. Strategies for isolation of all these genes are essentially identical. The use of restriction fragment length polymorphisms (RFLP's), of cytologically and morphologically defined deletions and of analysis of large DNA fragments (up to 10^6 bp) by pulsed field gel electrophoresis have proved critical in delineating regions of DNA that must contain the affected gene(s). The size of such regions, defined by any of these above techniques, ranges from $1\text{-}10\times10^6$ bp and there is a considerable technical challenge in physically isolating all of such a region. The immediate goal of such experiments must be to isolate the gene of interest. However, it is clear that a map of clones of DNA that correspond to such a large amount of DNA can be used to approach other, very different questions. In particular, the relationship of DNA sequence to chromosome structure and the higher order structure of DNA sequence (relationship of repeat sequence, organization of genes, and simple sequences) become accessible to direct analysis.

The technical challenge associated with cloning a large region of human DNA is the problem of scale. It is possible to isolate 20,000 or 40,000 bp of DNA in phage lambda or cosmid vectors. The latter vectors are most attractive since a single analysis or isolation of a cosmid clone provides information on twice as much DNA as a similar experiment on a phage clone. Cosmids have always proved difficult to make and handle. Many of these problems are inherent to working

with high molecular weight DNA. However, there are many reports of cosmids that show highly irreproducible growth characteristics. Variation in copy number (or yield) of DNA, instability and overgrowth of libraries by small, vector-sized molecules have all been observed. Little reported that the purpose of his group's current work has been an attempt to understand the sources of such variation and to reduce their effect by careful analysis of cosmid DNA replication systems.

Little and coworkers have found that variation in yield of cosmids is commonly due to transcriptional interference of the origin of replication by adventitious prokaryotic promoters in the human DNA, and that these interferences are remarkably common. These investigators have reduced this in two ways: (1) by the use of the phage lambda origin of replication, which is inherently less sensitive than other means to transcriptional interference, and (2) by the introduction of transcriptional terminators flanking the cloning sites.

Stability of clones (copy number control) was also investigated by Little and his group. Most conventional cosmids control copy number in cells by controlling the mass of cosmid DNA in the bacteria. A 5-kilobase (kb) vector will have 200 copies per cell but a 50-kb cosmid, based upon the same origin, will have 20 copies per cell, and small molecules will therefore be inherently more likely to survive in a population of growing *E. coli* than large ones. The alternative control method is copy number control--i.e., the cosmid replicates to a constant number of copies per cell, irrespective of size. Cosmids derived from Col E1 show predominantly mass control while the lambda cosmids are predominantly controlled by copy number and are therefore more stable.

Little and his group also studied the effects of cosmids upon bacterial growth rates. About 2 to 5 percent of human cosmid clones cause *E. coli* to grow slowly. They form tiny colonies on rich agar plates and do not reach saturation in similar liquid media within 12 to 14 hours. This effect seems to be caused by specific sequences contained within human DNA. The most likely candidate DNA sequences, according to Little, are chance homologies to RNA polymerase binding sites, operators, and, possibly, production of proteins toxic to *E. coli*. Little stated that not much can be done to alleviate this problem other than to reduce the copy number of the cosmid in the cell since a high copy number is most likely to generate adverse effects by binding a significant proportion of cellular repressor or RNA polymerase. All three "sequence specific" effects will be alleviated by the use of low copy numbers (1-2 copies per cell) vectors.

Another topic Little investigated was the question as to whether all sequences are clonable. Several groups have shown that DNA sequences repeated in a head to tail orientation in both phages and plasmids are highly unstable. The occurrence of these structures in human lambda phage clones is low (1 to 2 percent) but can be much more frequent in other organisms. The genetic background to support the growth of such sequences in phages (rec BC, sbc B) does not allow plasmid growth. More recently derived rec D rec A strains seem more successful but Little, as yet, has no data in what proportion of these clones this will be able to be propagated successfully.

Detailed analysis of large numbers of nematode DNA clones made in Col E1 and lambda ori cosmids show that cloning of sequences is not random and that many sequences are underrepresented in these libraries. One case of overrepresentation (ribosomal RNA genes) has been documented but the basis of this is not clear because such overrepresentation is rare.

The new generation of cosmid vectors has made some contribution to the ultimate goal of cloning all of a large region of DNA. According to Little, no single vector is ideal for this and he proposes that three different vector systems should be employed for full genomic coverage: high copy number lambda ori plasmids, low copy number (but inducible) cosmids, and phage lambda vectors. Two out of three of these systems are now

available, and mapping information will provide the necessary basis for a detailed, statistically significant description of clone distribution and coverage.

Molecular Techniques in Mammalian Genetics: Techniques for Chromosome Walking

Vector and cloning techniques have been developed to allow the molecular analysis of genetic distances in mammals--an essential step in the identification of genes defined by genetic mutations. Cloning techniques to simplify chromosome walking, to clone the ends of large fragments, to clone sequences containing rare description fragment sites as well as approaches to identify overlapping cosmids were described by H. Lehrach (European Molecular Biology Laboratory, Heidelberg, West Germany).

The last few years have seen the emergence of a number of new genetic and molecular approaches to an analysis of the human genome--developments, which have radically changed our prospects of ultimately understanding much of the information contained in it. Genetic analysis has improved dramatically with the use of DNA probes as genetic markers. This has allowed considerable progress in the development of a genetic map of the human genome as well as the identification of DNA markers closely linked to human mutations. Such markers are very important as diagnostic tools in genetic counseling or prenatal detection of the mutations, and open the way for a possible identification of the gene responsible for the mutation by a combination of molecular and genetic analysis steps. Similarly, the establishment of linkage maps must be combined with or complemented by the application of molecular techniques which allow the establishment of physical maps of chromosome regions or chromosomes. In both cases, however, the application of molecular cloning or DNA analysis is complicated by the disparity between the genetic distances of thousands of kilobase pairs (kbp), which typically will have to be analyzed, and the maximal amount of DNA (usually tens of kbp) which can be cloned in available vectors or easily analyzed by standard gel electrophoresis techniques. This gap of a factor of 10 to 100 between distances easily covered by the standard approaches of molecular biology and genetic analysis is now, however, being bridged by adapting both gel electrophoresis systems and cloning strategies to the analysis of hundreds to thousands of kbp (Poustka and Lehrach, 1986: Smith et al, 1986). In attempts to reach the environment of a mutation from a linked marker gene, Lehrach said that a series of steps will have to be taken. Such a stepwise analysis will become very inefficient or completely unfeasable for the analysis of larger regions of the genome or entire chromosomes. In this case, techniques allowing the parallel analysis of the region will have to be used.

Pulsed Field Gradient Gel Electrophoresis. The size separation of standard DNA gel electrophoresis systems is limited because large DNA molecules migrate independent of their molecular weight. This effect can be overcome by the use of pulsed field gradient gel electrophoresis (Schwartz and Cantor, 1984; Smith et al, 1986) or field inversion gel electrophoresis (Carle et al., 1986) in combination with enzymes cutting rarely in mammalian DNA. Systems of this type allow the demonstration of physical linkage of genetically closely linked DNA fragments and, using either information from single and double digest patterns or partial restriction mapping procedures, the establishment of restriction maps covering regions of millions of bp. Such maps, according to Lehrach, can be of major value in determining exact distances between probes; in localizing translocations, deletions, or insertions associated with a mutant phenotype; in localizing CpG-rich sequences often associated with the promoter sequence of housekeeping genes; and in complementing the genetic analysis techniques. Using these techniques, Lehrach and his group have analyzed regions from the environment of human and mouse mutations, especially the regions around the Duchenne Muscular Dystrophy locus, regions in the vicinity of the Huntington's Chorea and Cystic Fibrosis mutations in human, and

an area close to the Brachyury mutations in mouse.

Jumping Libraries. Chromosome walking (the repeated use of fragments from the end of a previously isolated clone to isolate clones extending further) has been used successfully in walking across genetic distances in organisms like *Drosophilia melanogaster* (fruitfly). This technique is much less successful in mammalian DNA. This is mainly due to the large number of steps which would have to be taken to cover distances of hundreds or thousands of kbp generally separating markers from mutations. To allow steps larger than the limited capacity of the available cloning vectors, the construction of chromosome jumping libraries is based on the deletion of all but the ends of large DNA fragments by a series of manipulations carried out before the cloning stage (Figure 1). Lehrach and his group have especially concentrated on chromosome jumping libraries prepared from DNA digested to completion with enzymes cutting rarely in mammalian DNA. This reduces the complexity of the libraries to be constructed and screened and simplifies the analysis of the resulting clones. Human jumping libraries have been constructed in Lehrach's laboratory using the enzyme Not I, Mlu I, BSSH II, and a combination of complete digestion with Not I and partial digestion with BSSH II. Libraries from mouse DNA have been constructed using the enzymes Not I, Mlu I, and BSSH II.

Linking Libraries. Sequences containing rare restriction sites can be selectively cloned out in the construction of "linking clone libraries" and constitute a well-defined subset of the entire genome (Figure 2). One obvious application is the use of such sequences to link up pairs of jumping clones. The identification of a Not I linking clone can therefore serve as an intermediate stop between two Not I chromosome jumps. In addition, different strategies to order linking clones relative to each other can be used, allowing the analysis of regions of the genomes in parallel. These can be based on identifying, for example, possible neighboring Not I linking clones

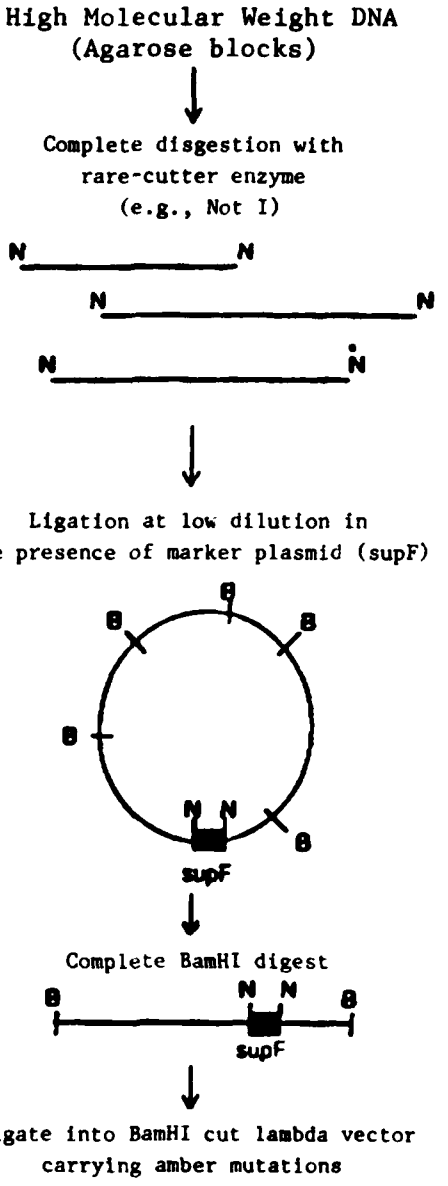


Figure 1. Schematic outline of the construction of jumping libraries N=Not I; B=BamHI.

by their expected hybridization to identical Not I bands in pulsed field gel electrophoresis (PFG) Southern blots. Alternatively, or in addition, fingerprinting techniques can be used to identify adjacent linking clones by their shared jumping clones. Lehrach and his group are testing some of these possibilities, using linking clones from the area of the Huntington's Chorea gene.

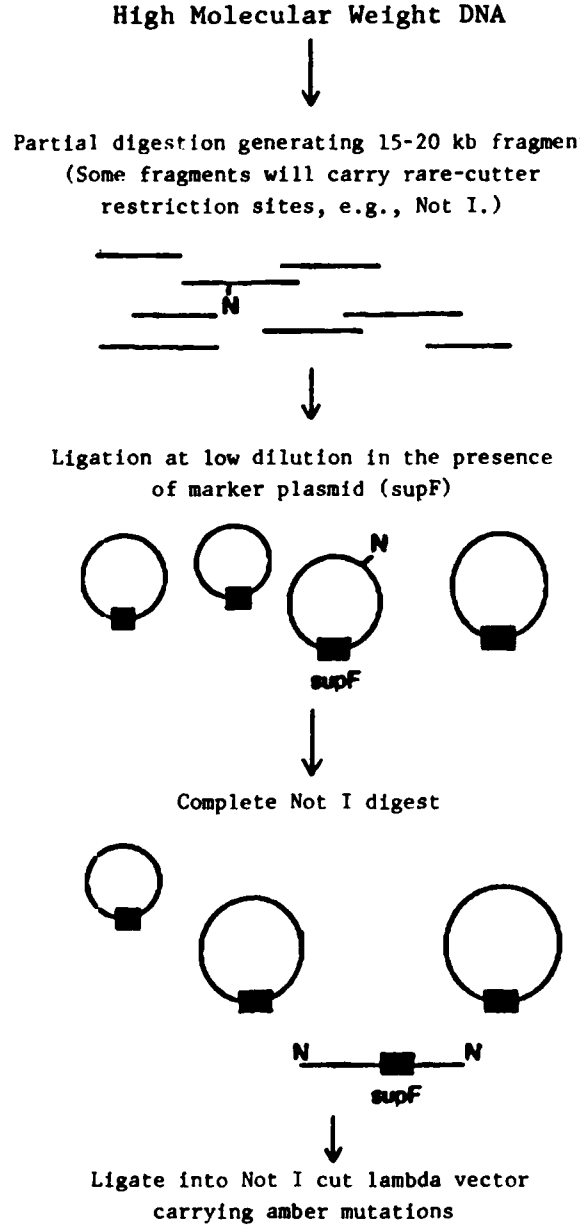


Figure 2. Schematic outline of the construction of jumping libraries N=Not I; B=BamHI.

Cloning of Fragments and Fragment Ends from PFG Gel Size Fractions. An alternative strategy to isolate clones from an area around an available marker clone or to clone sequences adjacent to rare cutting sites within such an area is the cloning of DNA fragments from a PFG size fractionation. In general, gel electrophoresis can only be expected to provide enrichments of 10- to at most 100-fold. Therefore, Lehrach provides a second enrichment by using somatic cell hybrids in which, for example, single human chromosomes or chromosome subregions have been transferred into a hamster or mouse cell line, allowing the differentiation of human and hamster or mouse clones by hybridization with total human DNA or cloned human repetitive sequences. Lehrach and his group have started such experiments to generate clones in the Duchenne Muscular Dystrophy region as well as the regions of the Huntington's Chorea and Cystic Fibrosis Loci.

Chromosome Linkup Techniques. The techniques described above will either provide a series of spaced marker fragments or, in the case of fragment cloning, reasonably complete coverage of a shorter segment. Strategies for complete coverage of larger regions of chromosomes or the entire human genome by ordered lambda or cosmid libraries will become more important as the molecular investigation moves from the analysis of few, easily accessible, well-defined, and medically highly important mutations to the analysis of more genes and more complex phenomena. Such ordered clone libraries would be of major importance in providing a high-resolution map of chromosomes, extending the genetic map down to a resolution of kb pairs (in the analysis of the genetic information contained in genetically defined regions) and serve as an essential first step in an ultimate analysis of regions of the genome by DNA sequencing.

Lehrach stated that while approaches to generate ordered libraries of small genomes have been tested, their application to mammalian chromosomes or the human genome will be extremely work intensive, leaving a major incentive to consider or develop potentially more efficient strategies. In the currently most efficient approach, DNA from random cosmids is cleaved by one enzyme (cutting approximately 10 times in, for example, the cosmid Hind III), the ends are radioactively labelled, and the DNA is

recleaved with a second enzyme generating short DNA fragments. These fragments are separated on sequencing gels and autoradiographed, and information about these radioactive fragments are stored in a data bank such that overlapping cosmids can be identified by common restriction fragments. A similar strategy, based on the identification of shared restriction fragments after agarose gel electrophoresis, is being applied to the analysis of the yeast genome. However, both techniques, without automation, are likely to be difficult to apply to mammalian genomes due to the very large number of clones (in the order of 1 million cosmids) to be handled individually. Therefore, Lehrach and his group have attempted to develop strategies to generate ordered clone libraries which do not require or at least minimize the individual manipulation of clones.

In Lehrach's approach, the identification of overlapping clones relies on shared hybridization (or nonhybridization) to a fairly large number of oligonucleotide probes selected to hybridize randomly with approximately one in three cosmids. Similar to the detection of genetic linkage between markers using recombinant inbred strains (Silver and Buckler, 1986), shared hybridization to single probes gives very little information, while very similar hybridization of many probes (analogous to similar strain distribution patterns) will constitute an overwhelming case for detection of an overlap (analogous to detection of linkage). Lehrach considers this approach to be an attractive one since the investigator can carry out the hybridization experiments on any number of DNA's from cosmid or lambda clones distributed as a regular array on membrane filters.

Lehrach then described a possible experimental setup which might, however, have to be modified to solve some experimental difficulties. DNA from cosmids or phage clones is transferred to nylon membranes and fixed by UV-crosslinking. The filters are then hybridized with oligonucleotide probes which have been end-labeled with ^{32}P and isolated from polyacrylamide gels. Hybridization/washing conditions are designed to reduce the effect of GC content on the hybrid stability. After washing, hybridizing colonies are detected by autoradiography and the hybridizing cosmids are identified by scanning the autoradiograph. Filters are then stripped of the old probes and the cycle is repeated, until 60 to 100 different probes had been scored (see below).

The length and sequence of the oligonucleotide probes are chosen to give hybridization probabilities of 10 to 50 percent to each cosmid. To achieve this hybridization probability with longer oligonucleotides and to reduce signals due to hybridization with bacterial DNA (a problem in attempts to use colony lifts), sequences overrepresented in mammalian DNA (and underrepresented in bacterial DNA) have been sought, taking advantage of the known nearest neighbor frequencies. For analysis, the hybridization pattern of each cosmid is stored as a binary number (reflecting hybridization versus no hybridization) with each digit corresponding to a particular probe. Overlapping clones will have similar binary numbers (with significantly higher coincidence than expected randomly). Additional information on the order of the oligonucleotide probes along the genome is provided by the loss (or acquisition) of probes by the linked cosmids (or phage). To test the feasibility of such an approach, Lehrach has carried out computer simulations using different numbers of cosmids distributed randomly over 1000 kbp. The order of the cosmids was randomized and reassortment attempted from the hypothetical hybridization patterns. This generates a number of curves showing the effect of varying the number of probes, hybridization probability, degree of overlap (which is proportional to the number of cosmids covering the 100-kbp test sequence) and the amount of error in scoring hybridization (or the effect of heterozygosity). Using this analysis, a hybridization probability of 30 percent with 8- to 12-fold genome coverage appears sufficient. With this combination, 60 to 80 probes are required to give a reasonable linkup. Error rates

of up to 5 percent are acceptable and can be compensated quite satisfactorally by a moderate increase in the number of probes used. Testing of the practicability of these experiment have shown that the short oligomers can be hybridized specifically to cosmid DNA. Difficulties, however, persist in the use of direct colony lifts of cosmid clones in hybridizations--a modification which would be clearly advantageous. In contrast, phage lifts seem to be quite acceptable for use in hybridizations. Further work in solving technical problems associated with handling large numbers of clones and in the development of modified colony hybridization protocols allowing the direct use of colony lifts in hybridization is in progress in Lehrach's laboratory.

3 EXPLOITING *E. COLI*

The Synthesis of Fusion Proteins in *E. Coli* and Their Uses in Raising Antibodies: From Gene to Antigen

Open reading frame expression vectors permit the isolation of monoclonal and polyclonal antibodies to defined domains of protein antigens and can also be used in direct functional studies. The enchanced immunogenicity of fusion proteins can be exploited to isolate antibodies of novel specificity. The same fusion protein constructions also allow accurate epitope mapping of existing antiprotein antibodies. These topics were discussed by D.P. Lane, Imperial Cancer Research Fund, Clare Hall Laboratories, London, UK.

Identification of an Open Reading Frame (ORF). Any newly acquired DNA sequence should be translated using the conventional and organism- and organelle-specific genetic codes in all three reading frames in both orientations using the appropriate DNA analysis software. The identification of an orf then depends on the imposition of an empirical set of criteria concerned with:

- Uninterrupted length of coding sequence
- Codon usage
- Identification of transcriptional start signals at the putative 5' end
- Identification of poly A^+ addition signal sites at the putative 3' end
- Identification of potential splice donor sites and acceptor sites flanking the orf.

Of all these criteria the first is the most easily spotted and least ambiguous. Any uninterrupted stretch of 100 or more amino acids should then be examined in the context of the remaining criteria. These are, of course, dependent on the organism from which the DNA has been isolated.

Expression of Orfs To Produce Antigen. Having identified a protative orf, the next step is to use this information to prepare a suitable antigen so that specific antibodies can be raised to part or all of the predicted amino acid sequence. There are two strategies that may be adopted at this point: peptide synthesis and expression in *E. coli*. Lane stated that complete reliance should not be placed on peptide synthesis although the two techniques are quite complementary.

Antibodies raised against short synthetic peptides coupled to large protein carrier molecules will sometimes also react with the same peptide sequence contained within a complete native protein. Thus, a short sequence can be selected from the orf chemically synthesized and used to induce antibodies to the orf. The two major advantages to this approach are the relative ease of preparation of the immunogen in large amounts and the site specificity of the resulting antibodies. Against this can be weighed a number of factors. Of greatest importance is its unreliability in that many peptides selected fail to induce antibodies that react with the orf. Some useful rules have emerged to generate peptide selection but these are not definitive according to Lane. Peptide synthesis is still a fairly expressive procedure, and peptides of greater than 20 to 30 amino acids in length cannot be routinely produced in adequate yield. This discourages most investigators from using peptide synthesis to scan a large fraction of the orf.

Lane also described orf expression vectors. In this method, sections of the DNA encoding the orf are cloned in the appropriate reading frame into *E. coli* vectors that are designed to lead to abundant expression of the orf as a fusion protein with a well-characterized *E. coli* protein. The vector is then introduced into an appropriate host strain and bacteria harboring the recombinant vector identified and used as a source from which to purify the fusion protein for use as an immunogen. The approach requires a series of vectors that permit the facile construction of the appropriate recombinants, and the easy isolation of the remaining fusion protein. The bacterial fusion protein partner should also confer a high degree of immunogenicity on the inserted orf. Given a vector that satisfies these criteria, then the approach has a number of compelling advantages. Primarily because large sections of the orf can be expressed, the technique has a very high success rate. The source of antigen is cheap and effectively limitless. It then becomes a straightforward step to express the selected orf fragment with alternate fusion partners in both prokaryotic and eukaryotic cells for direct functional studies.

Lane gave an example involving analysis of the SV40 large T orf using the pUR, pUC, and pSEM Cat_R1 vectors. Lane and his group have made extensive use of the pUR series of vectors. In the last 3 years they have constructed and analyzed about 40 different fusion proteins in these vectors. The orf's expressed have numerous viral and cellular genes ranging from yeast genes involved in splicing reactions to polypeptide hormone precursors. Their most detailed study has involved the 708 amino acid paparavirus oncogene, Large T. Lane presented these studies as a test case as he found that the results obtained with T reflect closely those seen with the other orfs studied. Figure 3 shows the structure of the pUR vectors. The key features are the multiple plasmid unique restriction sites available in all three reading frames into which the orf may be inserted. These sites are contained

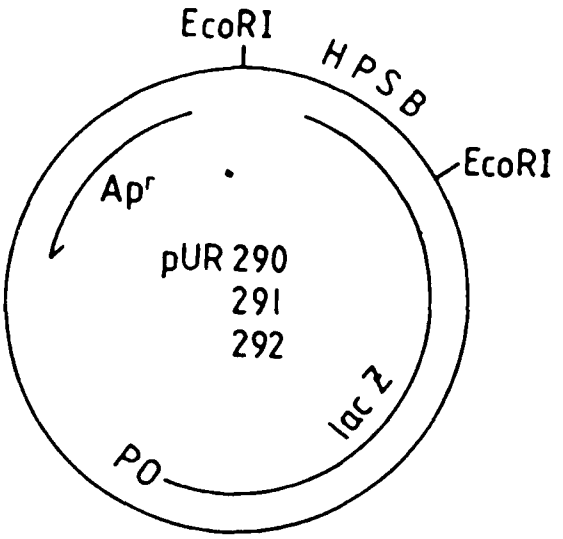


Figure 3. Structure of the pUR vectors.

within an "ECO R1 cassette" allowing ready excision of the insert and subsequent transfer to alternate vectors. Insertion into pUR results in the orf being placed at the extreme carboxy terminus of the β-galactosidase in the presence of inducers. Yet the system is almost completely repressed in the absence of inducers in i_q strains. Analysis of the resulting fusion proteins by sodium dodecyl sulfate (SDS) gel electrophoresis showed that they varied considerably with respect to the stability of the full-length product. The Hind III D fragment of T encoding amino acids 272 to 447 was completely stable while the Hind III A fragment fusion protein (amino acids 447 to 708) was extensively degraded. The high level of overproduction, however, has made it possible to isolate sufficient amounts of even the most sensitive proteins for immunization purposes. An extensive analysis of the reactivity of over 50 monoclonal antibodies with the fusion proteins has yielded a number of important results:

1. The fusion proteins can adopt native antigenic structure. All monoclonal anti-T antibodies raised to the native antigen will react with one or other of the fusion proteins even though

many of the antibodies will not recognize T that has been exposed to denaturants. Many of these antibodies raised to sensitive sites on the antigen will, however, only react with large insert-containing constructs; for instance, the antibody PAb 203 will react with the 272 to 708 fragment but not with either of the two fragments 272 to 447 or 447 to 708.

2. Antigenic silence. No monoclonal antibodies were found to react with the middle section of the antigen while a great number of antibodies were found to react with 5 percent of the protein carboxy terminal. This dominance of certain epitopes has been described for other antigens. Importantly, immunization of animals with fusion proteins that encompassed this silent region but lacked the dominant epitopes elicited a strong anti-T response and permitted the isolation of antibodies to the silent region (PAb 250 and PAb 251).

3. Shuttling to other vectors. The ECO R1 cassette can be shuttled into the unique R1 site within the Cat gene of the pSV_2 Cat derivative, pSEM Cat_R1. This results in an in-frame insertion at amino acid 72 of the Cat protein and presents the expression of the Cat fusion protein in both prokaryotic and eukaryotic cells. Lane has found that these Cat fusion proteins can retain orf-determined biological activities including sequence-specific DNA binding.

The Purification of Foreign Polypeptides Expressed in *E. Coli*

Gene cloning and expression in *E. coli* can provide an abundant source of eukaryotic peptides whose use is limited by low, natural availability. The polypeptides can be located in the cytoplasm of *E. coli* or secreted through the cell membrane. However, the mode of expression affects both the efficiency of production and the nature of the polypeptides themselves. This has implications for the purification techniques which must be developed. These topics were discussed by F. Marston (Celltech Ltd., Slough, UK).

Expression Vectors. A high level of accumulation of a "foreign product" in the host organism is often dependent on initiation of transcription by a strong promoter and the presence of multiple copies of the heterologous gene in each cell. It is of central importance, however, to have controlled expression from plasmids which are retained by organisms through successive generations. This is because the synthesis of recombinant products is a metabolic burden to the cells; cells which lose their plasmids soon outgrow plasmid-containing cells during a fermentation. Tight control of expression allows cells to be grown first to high biomass in the absence of heterologous gene expression, followed by growth for a limited number of generations with heterologous gene expression switched on.

The earliest expression plasmids used the strong promoters *trp* or *lac* to direct transcription. However, expression from such plasmids cannot be tightly controlled and they are unstable. Better control is achieved by using temperature-controlled expression plasmids incorporating promoters such as P_R and P_L obtained from the bacteriophage lambda (λ). These promoters can be controlled by the temperature-sensitive λ repressor gene product CI_{857} which, itself, is denatured at temperatures >37°C. Another feature of this vector is that it has two origins of replication which improve the efficiency of product accumulation by allowing control of plasmid copy number. One is a low copy number origin and the other a high copy number origin controlled by the λ P_R promoter. At 34°C or below, plasmids replicate from the low copy number origin, and replication from the other origin is controlled by the CI_{857} repressor. During this phase, growth to high biomass can be achieved before the fermentation temperature is increased from 38°C to 42°C, at which point the CI_{857} repressor is denatured, plasmid copy number is amplified, and heterologous gene expression is switched on.

Modes of Expression. Gene construction can be used to determine the location of the foreign proteins. The proteins may either be located in the cell cytoplasm or, by incorporating a leader sequence before the coding sequence,

products may be secreted through the cell membrane. There are two general strategies for the synthesis of proteins which are located intracellularly. The gene can be cloned in a frame with synthetic or bacterial coding sequences and expressed as a fusion protein; alternatively, the foreign gene is expressed directly. Recombinant products accumulate to greater levels when expressed intracellularly (up to 25 percent of total cell protein) than when they are secreted (up to 1 percent of total cell protein). However, many heterologous polypeptides located in the cytoplasm are insoluble and require specific solubilization techniques to yield active products for purification.

Inclusion Bodies. Insoluble recombinant proteins accumulate in *E. coli* in the discrete form of inclusion bodies as first reported in 1982 for proinsulin. Phase contrast microscopy shows inclusion bodies to be large, in relation to the size of the cells, highly refractile, and often located at the poles of the cells. No distinct structure within the inclusion bodies is revealed by electron microscopy and they are apparently not in contact with the cell membrane. The existence of the insoluble polypeptides in the form of inclusion bodies is useful for purification as they are dense and sediment readily during low-speed centrifugation. Under these conditions, the inclusion bodies are purified away from a large proportion of the cell debris, as well as from the soluble proteins. Further purification can be achieved by washing the isolated inclusion bodies with detergent or urea to recover active soluble protein. The isolated inclusion bodies must be denatured and unfolded. Conditions must then be adjusted to allow the polypeptides to refold into the correct conformation.

Fusion Proteins. Small polypeptides, when expressed in *E. coli* are often degraded rapidly. This can be prevented by expressing the eukaryotic genes as fusions with partial or entire bacterial coding sequence. There are many examples of fusion proteins which are in an insoluble form when synthesized in *E. coli* such as somatostatin, human calcitonin, urogastrone, β-globin, bovine growth hormone, etc. The fusion proteins contained in inclusion bodies (isolated as described above) must be solubilized before further purification is possible and this has been achieved by the use of detergents and denaturants. These solubilization agents were maintained in the buffers used for column chromatography. Purified fusion proteins have been used in diagnostic development and analytical research but, in general, the foreign polypeptide is required free from the bacterial sequence. The strategy used to achieve this is to place a cleavage site between the C-terminus of the prokaryotic sequence and the N-terminus of the eukaryotic coding sequence. A unique cleavage site, not present in the sequence of the recombinant protein is the ideal arrangement so that the foreign protein itself is not cleaved. There are two approaches for cleavage to be effected, either chemical or enzymatic. Solubilization may be required before the cleavage reaction can be performed. This is illustrated in the protocol used to cleave an insoluble trp LE-bovine growth hormone fusion protein. An acid labile-asp-pro-cleavage site was engineered into this fusion protein and isolated inclusion bodies were incubated in 6 Molar guanidine hydrochloride at low pH to release the bovine growth hormone.

When enzymatic cleavage is used, conditions for solubalization are required in which the enzyme is still active. An insoluble chloramphenicol acetyl transferase-lys-arg-calcitonin fusion protein was cleavable because the enzyme used (Clostripain) is stable in up to 6 M urea. In contrast, after solubilization in denaturant, dilution may be required before addition of enzyme, as found for cleavage of a trp E-lys-epidermal growth factor fusion protein by endoproteinase Lys C. The longer the recognized cleavage site, the less likely it is to be found in the coding sequence of the recombinant gene. This was the approach taken in the construction of trp E-epidermal growth factor fusions with a collagenase cleavage site. If the

cleavage site is not unique, then internal lysis sites can be protected during cleavage. For example, with a β-galactosidase-β-endorphin fusion, internal lysine residues were reversibly blocked by citraconylation and trypsin was used to digest the lys-arg-residues immediately before the N-terminus of the foreign polypeptide.

Purification of Fusion Proteins. Fusion proteins can be constructed to facilitate purification by using bacterial or synthetic nucleic acid sequences coding for peptides. One approach was to fuse a sequence coding for polyarginine to the 3' end of the epidermal growth factor (urogastrone) gene. Cation exchange chromatography used to purify the fusion protein, was particularly effective, since most bacterial proteins are acidic and therefore negatively charged at the pH of 5.5, which was used. Carboxypeptidase B was used to digest the polyarginine tail, and further cation exchange chromatography yielded highly purified hormone. There are examples of fusion proteins which are soluble in *E. coli* such as λcII-α-antitrypsin and the β-galactosidase-pre S surface antigen protein of hepatitis B virus. The latter was constructed as a purification fusion. Affinity chromatography on p-aminophenyl-β-D-thiogalactoside-Sepharose yielded fusion protein >90 percent pure.

Direct Expression. There are examples of heterologous proteins produced in *E. coli* in active, soluble forms (α_1 anti-trypsin, interleukin-2, human lymphotoxin, interfeon-α, etc.) but there are few obvious common features to explain why these proteins are soluble. Immunopurification was a major step in the purification of lymphotoxin and AIDS virus p24 gag protein expressed in *E. coli*. From the literature, it is apparent that most eukaryotic polypeptides synthesized by direct expression in *E. coli* are insoluble. The conditions which will successfully solubilize recombinant proteins from inclusion bovines are those which *in vitro* will denature native proteins. These agents (urea, guanadine-hydrochloride, acetonitrile, propanol, etc.) disrupt different types of noncovalent focus such as hydrogen bonds, ionic and hydrophobic interactions. The effectiveness of a particular agent differs between proteins. Thus, the nature of the interaction within inclusion bodies differs and is protein specific. Further support for these conclusions is given by the fact that specific conditions such as pH, temperature, time, and ionic environment have been defined for the solubilization of each protein. Covalent interactions, such as disulphide bonds, may also exist in inclusion bodies since thiol reagents are required to solubilize some recombinant proteins--for example, bovine growth hormone and IgG light chain. Disulphide bonds are unlikely to have formed in the reduced environment of the *E. coli* cytoplasm and probably form in air during lysis.

Refolding. After solubilization from inclusion bodies and cleavage for fusion proteins, the polypeptides must be refolded to yield active protein. Refolding is achieved by buffer exchange out of denaturing conditions using, for example, dialysis or dilution. Small polypeptide hormones such as β-endorphin and calcitonin apparently fold spontaneously. For larger polypeptides, specific conditions must be defined in order to obtain correctly folded protein, according to Marston. A key factor influencing the recovery of active product is protein concentration. This must be low enough to allow intramolecular interactions to occur in preference to intermolecular interactions.

Marston presented some examples of refolding protocols for insulin, β-globin, and prochymosin. Insulin A chain and insulin B chain have been cloned separately in *E. coli* and each has a β-galactosidase fusion protein. Isolated inclusion bodies were solubilized in 6-M guanidine HCl and 1-percent 2-mercaptoethanol which caused at least some intermolecular disulphide bonds to be reduced. After dialysis out of denaturant, the fusion proteins were cleared using cyanogen bromide in 70-percent formic acid. The pellet obtained after rotary evaporation was dissolved in 8-M guanidine HCl and S-sulphonated at p H9. After 24 hours

the pH was adjusted to 5 and the solution clarified by centrifugation. To reconstitute insulin, S-sulphonated A and B chains were mixed, lyophilized, and incubated first at pH 4.5 in the presence of 2-mercaptoethanol and then at pH 9.6 to 10.6. This reconstitution protocol yielded 10- to 15-percent correctly folded insulin as judged by radioimmunoassay.

λcIII-β globin is another example of an insoluble fusion protein which has been successfully denatured and refolded. Inclusion bodies were isolated, washed in Triton X-100 and solubilized in 8-M urea, buffered at pH 5, containing EDTA and dithiothreitol. The fusion protein was purified by ion-exchange chromatography and gel filtration in buffers containing denaturant, EDTA, and dithiothreitol. β-globin was released from the fusion by digestion with factor Xa. The protein was then dissolved to a concentration of 5 mg/mL in 8 M urea, containing dithiothreitol and buffered at pH 8. λcIII-β globin refold occurred by dilution to 0.3 mg protein/mL and reconstituted with cyanoheme and α-chain in 1:2 molar excess. The oxygen-binding properties of the recombinant hemoglobin were essentially the same as those of authentic hemoglobin.

Prochymosin is a directly expressed protein which has been refolded to yield biologically active protein by Marston and her group at Celltech. Isolated, washed inclusion bodies were solubilized in 8-M urea and refolding--effected in two stages. The urea extract was diluted into an alkaline buffer and the pH maintained at 10.7. After a period of incubation the pH was adjusted to 8.0 and the prochymosin remained in solution. The yield of active enzyme was found to depend critically on the level of dilution and pH of the diluent. This process routinely produced a yield of 25 to 50 percent of the expressed enzyme in a soluble, active form. Both the urea and alkali steps were required. Neither step alone produced an equivalent yield. After refolding, the prochymosin was 30-percent pure and ion-exchange chromatography was used to produce prochymosin >90-percent pure with an overall yeild of 22 percent.

Secretion. *E. coli* possesses two cell membranes, the cytoplasmic membrane and the outer membrane, which are separated by the periplasmic space. Proteins located in the periplasm or the outer membrane are synthesized in the cytoplasm and transported through the cytoplasmic membrane. Those proteins are synthesized as precursors, with an N-terminal signal sequence, which may be cleaved during secretion. Signal sequences are an absolute requirement for secretion and are always cleaved during the process of secretion. Secretion of foreign proteins from *E. coli* can overcome insolubility, but *E. coli* does not naturally secrete high levels of proteins, and expressed levels of secreted recombinant proteins have been low. Since few proteins are secreted from *E. coli*, less extensive purification of recombinant proteins is required than for proteins located in the cytoplasm. Epidermal growth factor and β-endorphin, secreted into the periplasm have been purified to apparent homogeneity in two steps, gel filtration followed by reverse phase high-pressure liquid chromotography. Human growth hormone secreted into the periplasm of *E. coli* has been purified to >90 percent homogeneity, also in a two-stage process using ion-exchange chromatography and gel filtration. The purified, secreted human growth hormone was characterized by assignment of disulfide bonds by circular dichroism spectroscopy. These data indicated that the secondary structure of recombinant human growth hormone was identical to that of authentic human growth hormone.

Characteristics of the *E. Coli* Expression System. There is concern over the use of *E. coli* as a production system for foreign proteins because of the presence of endotoxins or lipopolysaccharides (LPS) in the cell wall. LPS is pyrogenic and must be removed to yield a safe product, particularly if the protein is to be administered as a therapeutic agent. There are several chromatographic steps which can be used to remove LPS, some of which are commonly used in protein purification.

Eukaryotic polypeptides synthesized in *E. coli* can differ from the authentic molecules. Directly expressed molecules may possess an additional methionine residue at the N-terminus. *E. coli* does possess enzymes which catalyze the removal of the initiating methionine, but the efficiency with which this amino acid is cleaved from recombinant polypeptides is variable.

Soluble active products can be recovered from the insoluble foreign polypeptides synthesized by *E. coli*. However, there is the possibility that the solubilization and refolding processes result in modification of the polypeptides. There are reports of refolded foreign polypeptides which have been characterized by disulphide bond assignment, nuclear magnetic resonance and x-ray crystallography. No significant differences were detected between the authentic and recombinant proteins.

E. coli does not possess the enzymes for catalyzing post-translational modifications such as glycosylation, amidation and acetylation. Some eukaryotic polypeptides syntesized in *E. coli* have been found to be active despite the fact that they were not glycosylated. These include interferon-β, interferon-α and interleukin-2. *In vitro* systems have been developed to perform certain post-translational modifications of polypeptides isolated from *E. coli* such as amidation of calcitonin and acetylation of desacetyl thymosin.

Marston emphasized that there are advantages and disadvantages in using *E. coli* to express heterologous proteins. Large amounts of foreign proteins can be synthesized per liter of fermentation by intracellular expression, but the products are often insoluble. Secretion from *E. coli* overcomes the insolubility problem but expressed yields are low. If *E. coli* is to be used as a secretion system then further development is required.

Eukaryotic polypeptides synthesized by and purified from *E. coli* are now in therapeutic use--interleukin-2, interferon-α, β, and TNF (tissue necrosis factor) are currently in clinical trials while insulin and human growth hormone are in clinical use. Therefore, *E. coli* is a viable production system for the manufacture of therapeutic proteins.

The Protein Engineering of Subtilisin

The enzyme subtilisin is being used as a model system in several laboratories for protein engineering studies. Alterations in the enzyme have been made in its stability, substrate specificity, and pH optimum. From the basic information from the investigation, electrostatic effects in enzymes can be determined. The lecture on this topic was given by A. Fersht (Department of Chemistry, Imperial College of Science and Technology, London, UK).

Subtilisin is a serine protease which is secreted by many *bacilli*. The enzymes from *Bacillus licheniformis* and, to a lesser extent, from *Bacillus amyloliquefaciens* are used extensively in industrial processes. Vast amounts of the *licheniformis* enzyme are used in soap powder. The enzyme, containing some 275 amino acids, has a relatively broad specificity but is most specific for tyrosine in the primary binding site. It is a typical serine protease involving a catalytic triad of a nucleophilic serine hydroxyl (Ser-221) and a base (His-64) which is hydrogen bonded to an aspartate (Asp-32).

Subtilisin provides an excellent system for extensive protein engineering studies, both as a model and also for potential use in biotechnology. Its merits are the following: it is secreted at a very high level from the *Bacillus amyloliquefaciens* containing the cloned gene (about 1 g/L); its crystal structure has been solved at high resolution x-ray as have complexes with polypeptide inhibitors; and it may be readily and quantitatively assayed by steady-state kinetics and active site titration.

The gene has been cloned and expressed in a number of laboratories but published work is principally from the US groups at Genentech/Genecor and Genex and from the UK's Imperial College. The goals of the research have been to:

- Improve thermal stability

- Improve stability to oxidation by bleach
- Change substrate specificity
- Change pH profile
- Increase activity under conditions used in processes.

Improvement of Thermal Stability. Attempts have been made by both Genex and Genentech/Genecor to improve the thermal stability of the enzyme by engineering disulphide bridges which were predicted to be optimal by computation based on the known geometry of disulphide bonds in other proteins. However, no improvement in stability was found, partly because autolysis is a major problem. Fersht stated that it is clear we do not yet understand the energetics of stabilization of disulphide bridge formation.

Stability to Oxidation. Subtilisin in washing powder encounters bleach, which rapidly oxidizes sulphurs in methionine and cysteine residues. Genentech/Genecor have replaced methionine 222 by all other 19 amino acids using cassette mutagenesis and have found the enzyme still active for certain substrates, although at a lower rate. The modified enzymes are, however, much more resistant to oxidation by bleach.

Substrate Specificity. Genentech/Genecor have altered the specificity of the enzyme by systematically filling up the primary binding site with bulky side chains. Gly-166 in the cleft was the target for cassette mutagenesis. Bulky hydrophobic side chains were found to lower the activity to tyrosine-containing substrates, but to increase activity of the enzyme towards smaller side chains in a rational manner.

Tailoring the pH Dependence. Fersht and his group at Imperial College, London, have been able to raise and lower the pH optimum of the enzyme by making rational changes in surface charge. Negatively charged groups stabilize the protonated, inactive form of histidine-64 while positively charged groups stabilize the subtilisin enzyme. Thus, mutation of surface residues from negative to positive charge, systematically lowers the pKa governing activity and increases catalytic activity at low pH. Catalytic activity has been increased under certain conditions such as engineering substrate specificity and tailoring pH dependence.

Experiments aimed at the above goals have provided basic information on protein structure and activity. According to Fersht, much has been learned about binding, reactivity, and electrostatic effects which will be applicable to other systems. Fersht considers it likely that engineered subtilisin will replace the natural product for certain processes.

4 EXPRESSION IN MAMMALIAN CELL CULTURE AND TRANSGENIC MICE

Eukaryotic Expression Vectors

Genetic engineering provides the potential to produce, in large quantity, scarce or even completely novel proteins. However, it has become clear that prokaryotic expression systems are often inappropriate for complex eukaryotic proteins. The fermentation of animal cells transformed with suitable eukaryotic vectors can provide a viable alternative, as discussed by C. Hentschel (Celltech Limited, Slough, UK).

E. coli has been used successfully to synthesize a variety of foreign gene products. However, it has become increasingly clear that the proteins produced may not be identical to their natural eukaryotic counterparts. Eukaryotic proteins typically undergo a number of post-translational events which are potentially as important for their function as is the sequence of the polypeptide chain. These events, which include modifications such as glycosylation, phosphorylation, and amidation; proteolytic cleavages of precursor proteins; macromolecular assembly; and (often) secretion generally can not be performed in bacteria. The comparatively reducing environment in bacteria, such as *E. coli*, can also lead to the formation of incorrect disulphide bridges. The absence of appropriate modifications can profoundly affect the conformation and solubility of a protein, and for some proteins, correct secondary modifications may be absolutely required for the function of the molecule. For

instance, particular oligosacharide sequences are required by the glycohormone, chorionic gonadotrophin in order to bind to its receptor. Steric effects of glycosylation are also likely to be important in the production of vaccines, which often depend for their immunogenicity on carbohydrate compounds as well as on the precise tertiary structures of their polypeptides. Carbohydrates may similarly be of significance in the production of pharmaceuticals designed for repeated therapeutic use. In this case, it is desirable to minimize the risk of stimulating an immune response. Therefore, a tertiary structure as close as possible to the natural conformation is required.

The profound effect which differing expression systems can have on the final product is well illustrated by the case of immunoglobulin genes which have been cloned and expressed in a variety of different cell types as a first step towards obtaining novel antibodies by genetic engineering. In *E. coli*, expression of genes for immunoglobulin (IgG) heavy and light chains together in the same cell fails to yield any detectable *in vivo* antigen-binding activity. When expressed in yeast, functional association of the two polypeptide chains has been obtained but the hapten binding activity of the antibody produced is only about 0.5 percent that of the natural protein. On the other hand, expression in mouse myeloma cells yields completely functional antibodies. For some recombinant proteins, the choice of cell type is even more critical. Thus, for example, the gene for the clotting agent, Factor IX, can yield functional protein when transferred into baby hamster kidney (BHK) cells or rat hepatoma cells, but not when expressed in fibroblasts. These lack the enzymes required for carboxylation of glutamic acid residues, a modification known to be essential for the protein's activity.

Gene Transfer Methods for Eukaryotic Vectors. The evolution of eukaryotic expression vectors has relied on progress in the methodology of gene transfer in mammalian cells. A commonly used procedure exploits the fact that DNA, for example, cloned in a bacterial vector, can be taken up by the cells either as a coprecipitate with calcium phosphate or bound to DEAE-dextran. Once inside the nucleus, the DNA can become inserted, essentially at random, into one or more of the host chromosomes. Stable cell lines containing integrated DNA are generated at low frequencies so a selectable marker is usually included on the vector to eliminate transformed cells. An important and highly efficient alternative form of gene transfer can be provided by viral infection, for example, by retrovirus vectors.

Optimizing Expression from Integrated Vectors. Introduction of genomic DNA sequences into mammalian cells by one of the above techniques can be sufficient to lead to detectable expression of transferred genes, provided that essential control sequences are present in addition to the coding region. In some cases--for example, IgG heavy chain genes expressed in myeloma cells--high levels of protein secretion can be achieved. In others, the level of expression may be rather low, even for a naturally abundant protein expressed in a cell type capable of high-level expression. For instance, when globin genes are transferred into an erythroleukemia cell line, the level of globin messenger RNA (mRNA) produced per gene is 1/10 to 1/100 that of the endogenous genes. Similarly, transfected IgG light chains tend to be poorly expressed in myeloma cells.

One way to provide the signals specifying high-level expression is to insert the coding sequence into the genome of an animal virus at an appropriate location downstream of a strong promoter in place of one or more viral genes. Adenovirus vectors based on this approach have been used to express efficiently a number of genes. Retroviruses have similarly been used to express high levels of commercially useful proteins (hepatitis surface B antigen, growth hormone, tissue plasminogen activator, etc.).

An alternative form of vector construction combines functional elements obtained from different sources. In this case, a complete transcription unit is formed by providing the coding sequence

firstly with a suitable promoter-enhancer combination for efficient transcription in the particular cell type chosen. Secondly, appropriate, transcribed sequences 5' and 3' to the protein-coding may be provided to enhance translation and message stability. It may also be advantageous to coexpress a transacting transcriptional activator gene in the same cell. A number of genes have now been identified whose protein products specifically enhance transcription from particular promoters. Also, it has been shown recently that the efficiency of mRNA translation can sometimes be increased by transacting factors. For example, translation from the adenovirus tripartite leader can be dramatically increased by interaction with the 'virus associated' (VA) RNA's of adenovirus when these are coexpressed in the same cell.

Gene Amplification. The amount of protein product of transfected genes is often found to be roughly proportional to the number of functional copies of the gene present. Thus, a strategy which allows an increase in the copy number of the integrated genes is clearly desirable. The possibility of intentionally increasing the copy number of transfected genes stems from the discovery of the phenomenon of gene amplification. If cultured cells are subjected to sequentially increasing amounts of toxic drug, variant clones can frequently be selected which are more resistant to the drug than wild-type cells. In the majority of cases, this is due to the overproduction of an essential enzyme whose activity the drug is inhibiting. The overproduction of the enzyme has frequently been shown to be due to an increase in the copy number (i.e., amplification) of the structural gene coding for the enzyme. Over 10 cases of endogenous gene amplification have definitely been identified in mammalian cells. Examples such as dihydrofolate reductase (DHFR), asparagine synthetase, and metallothionein genes appear to have in common only the fact that they code for essential proteins for which an inhibitory drug is available. Furthermore, since the region of amplified DHFR is always much greater than just the selected structural gene (often more than 1000 kb is amplified intact) a nonselected, cointegrated gene can be coamplified. The potential of the system to overproduce a nonselected protein product was first demonstrated by the use of a complimentary DNA (cDNA) copy of a wild-type DHFR gene to transform mutant Chinese hamster ovary (CHO) cells lacking DHFR activity to a $DHFR^+$ phenotype. Clones were subsequently selected which were resistant to methotrexate, an inhibitor of DHFR, and hence amplification of the structural gene was obtained. The gene for the small t-antigen of SV40, which was present on the vector, was also amplified and large quantities of t-antigen were produced (about 10 percent of the total cell protein). Coamplification of an unrelated structural gene with a DHFR gene has since been used to produce several commercially useful proteins in large amounts and, in fact, the highest reported levels of expression have been achieved with this system (Hepatitis B surface antigen, interferon-β, tissue plasminogen activator).

Vectors Based on Viral Replicons. The alternative method of achieving persistence of the introduced gene in the host cell is to provide an origin of DNA replication as part of the vector. A particularly useful animal virus for the construction of expression vectors has been the bovine papillomavirus (BPV) which replicates episomally in rodent fibroblast cells. The copy number is relatively low (up to several hundred/cell) but the major advantage is that the virus does not kill the host, and hence stable cell lines can be produced. Vectors containing either the whole virus genome or certain subgenomic fragments induce oncogenic transformation so that cells grow as dense foci on a monolayer, and this phenotype can be used as a means of identifying transfected clones. In several cases, BPV vectors have been shown to remain episomal in the same way as the intact virus. However, it is also now clear that many BPV-based vectors can integrate into the host DNA, often as one or more head-to-tail tandem arrays. Either way, the relatively high copy number

instance, particular oligosacharide sequences are required by the glycohormone, chorionic gonadotrophin in order to bind to its receptor. Steric effects of glycosylation are also likely to be important in the production of vaccines, which often depend for their immunogenicity on carbohydrate compounds as well as on the precise tertiary structures of their polypeptides. Carbohydrates may similarly be of significance in the production of pharmaceuticals designed for repeated therapeutic use. In this case, it is desirable to minimize the risk of stimulating an immune response. Therefore, a tertiary structure as close as possible to the natural conformation is required.

The profound effect which differing expression systems can have on the final product is well illustrated by the case of immunoglobulin genes which have been cloned and expressed in a variety of different cell types as a first step towards obtaining novel antibodies by genetic engineering. In *E. coli*, expression of genes for immunoglobulin (IgG) heavy and light chains together in the same cell fails to yield any detectable *in vivo* antigen-binding activity. When expressed in yeast, functional association of the two polypeptide chains has been obtained but the hapten binding activity of the antibody produced is only about 0.5 percent that of the natural protein. On the other hand, expression in mouse myeloma cells yields completely functional antibodies. For some recombinant proteins, the choice of cell type is even more critical. Thus, for example, the gene for the clotting agent, Factor IX, can yield functional protein when transferred into baby hamster kidney (BHK) cells or rat hepatoma cells, but not when expressed in fibroblasts. These lack the enzymes required for carboxylation of glutamic acid residues, a modification known to be essential for the protein's activity.

Gene Transfer Methods for Eukaryotic Vectors. The evolution of eukaryotic expression vectors has relied on progress in the methodology of gene transfer in mammalian cells. A commonly used procedure exploits the fact that DNA, for example, cloned in a bacterial vector, can be taken up by the cells either as a coprecipitate with calcium phosphate or bound to DEAE-dextran. Once inside the nucleus, the DNA can become inserted, essentially at random, into one or more of the host chromosomes. Stable cell lines containing integrated DNA are generated at low frequencies so a selectable marker is usually included on the vector to eliminate transformed cells. An important and highly efficient alternative form of gene transfer can be provided by viral infection, for example, by retrovirus vectors.

Optimizing Expression from Integrated Vectors. Introduction of genomic DNA sequences into mammalian cells by one of the above techniques can be sufficient to lead to detectable expression of transferred genes, provided that essential control sequences are present in addition to the coding region. In some cases--for example, IgG heavy chain genes expressed in myeloma cells--high levels of protein secretion can be achieved. In others, the level of expression may be rather low, even for a naturally abundant protein expressed in a cell type capable of high-level expression. For instance, when globin genes are transferred into an erythroleukemia cell line, the level of globin messenger RNA (mRNA) produced per gene is 1/10 to 1/100 that of the endogenous genes. Similarly, transfected IgG light chains tend to be poorly expressed in myeloma cells.

One way to provide the signals specifying high-level expression is to insert the coding sequence into the genome of an animal virus at an appropriate location downstream of a strong promoter in place of one or more viral genes. Adenovirus vectors based on this approach have been used to express efficiently a number of genes. Retroviruses have similarly been used to express high levels of commercially useful proteins (hepatitis surface B antigen, growth hormone, tissue plasminogen activator, etc.).

An alternative form of vector construction combines functional elements obtained from different sources. In this case, a complete transcription unit is formed by providing the coding sequence

firstly with a suitable promoter-enhancer combination for efficient transcription in the particular cell type chosen. Secondly, appropriate, transcribed sequences 5' and 3' to the protein-coding may be provided to enhance translation and message stability. It may also be advantageous to coexpress a transacting transcriptional activator gene in the same cell. A number of genes have now been identified whose protein products specifically enhance transcription from particular promoters. Also, it has been shown recently that the efficiency of mRNA translation can sometimes be increased by transacting factors. For example, translation from the adenovirus tripartite leader can be dramatically increased by interaction with the 'virus associated' (VA) RNA's of adenovirus when these are coexpressed in the same cell.

Gene Amplification. The amount of protein product of transfected genes is often found to be roughly proportional to the number of functional copies of the gene present. Thus, a strategy which allows an increase in the copy number of the integrated genes is clearly desirable. The possibility of intentionally increasing the copy number of transfected genes stems from the discovery of the phenomenon of gene amplification. If cultured cells are subjected to sequentially increasing amounts of toxic drug, variant clones can frequently be selected which are more resistant to the drug than wild-type cells. In the majority of cases, this is due to the overproduction of an essential enzyme whose activity the drug is inhibiting. The overproduction of the enzyme has frequently been shown to be due to an increase in the copy number (i.e., amplification) of the structural gene coding for the enzyme. Over 10 cases of endogenous gene amplification have definitely been identified in mammalian cells. Examples such as dihydrofolate reductase (DHFR), asparagine synthetase, and metallothionein genes appear to have in common only the fact that they code for essential proteins for which an inhibitory drug is available. Furthermore, since the region of amplified DNA is always much greater than just the selected structural gene (often more than 1000 kb is amplified intact) a nonselected, cointegrated gene can be coamplified. The potential of the system to overproduce a nonselected protein product was first demonstrated by the use of a complimentary DNA (cDNA) copy of a wild-type DHFR gene to transform mutant Chinese hamster ovary (CHO) cells lacking DHFR activity to a $DHFR^+$ phenotype. Clones were subsequently selected which were resistant to methotrexate, an inhibitor of DHFR, and hence amplification of the structural gene was obtained. The gene for the small t-antigen of Sv40, which was present on the vector, was also amplified and large quantities of t-antigen were produced (about 10 percent of the total cell protein). Coamplification of an unrelated structural gene with a DHFR gene has since been used to produce several commercially useful proteins in large amounts and, in fact, the highest reported levels of expression have been achieved with this system (Hepatitis B surface antigen, interferon-β, tissue plasminogen activator).

Vectors Based on Viral Replicons. The alternative method of achieving persistence of the introduced gene in the host cell is to provide an origin of DNA replication as part of the vector. A particularly useful animal virus for the construction of expression vectors has been the bovine papillomavirus (BPV) which replicates episomally in rodent fibroblast cells. The copy number is relatively low (up to several hundred/cell) but the major advantage is that the virus does not kill the host, and hence stable cell lines can be produced. Vectors containing either the whole virus genome or certain subgenomic fragments induce oncogenic transformation so that cells grow as dense foci on a monolayer, and this phenotype can be used as a means of identifying transfected clones. In several cases, BPV vectors have been shown to remain episomal in the same way as the intact virus. However, it is also now clear that many BPV-based vectors can integrate into the host DNA, often as one or more head-to-tail tandem arrays. Either way, the relatively high copy number

has led to high level of expression in mouse fibroblasts for a number of genes such as Hepatitis B Surface antigen and tissue plasminogen activator (tPA).

Hentschel stated that there is no problem about scale-up production of eukaryotic recombinant proteins. There is already a great deal of experience in the bulk fermentation of mammalian cells for the production of proteins, chiefly for use as vaccines and of hybridoma cells for the manufacture of monoclonal antibodies. The maximal expression levels of greater than 3×10^8 molecules/cell/day attained for tPA, Hepatitis B surface antigens, and human growth hormone compare favorably with monoclonal antibody production; a good hybridoma producing about 7×10^8 molecules of antibody/cell/day. The techniques described above clearly allow previously scarce proteins to be produced in equally large quantities using similar fermentation technology.

The Production of Transgenic Mice by the Direct Microinjection of Cloned DNA's

The methods used to produce transgenic mice by the microinjection of cloned DNA into fertilized one-cell eggs were described by D. Murphy (Molecular Embroyology Laboratory, National Institute for Medical Research, London, UK) using illustrations from his laboratory, where the technique is being used to study the molecular genetic mechanisms of hormone gene expression and cancer.

The past decade has seen the molecular cloning and structural characterization of a large number of mammalian genes. The function and regulation of these genes has been studied by gene transfer experiments. Wild-type constructs and mutated derivatives have been transfected into tissue culture cells in order to identify *cis*-acting regulatory elements and to investigate the physiological consequences of the expression of the gene product. However, even if appropriate tissue culture systems exist for the gene of interest, only limited perspectives on gene expression can be derived from such *in vitro* experiments. Ultimately, gene function and expression must be studied within the complexities of the whole organism. A number of techniques have been developed that allow the introduction of defined DNA sequences into the germ line of mice and other mammals. Once inserted, these sequence--termed transgenes--are stably transmitted from generation to generation. Of fundamental importance is the observation that transgenes are often expressed, and subject to correct developmental, tissue specific, and physiological regulation. It is therefore now possible to analyze the role and regulation of specific cloned genes within the whole organism--the transgenic organism.

Routes to the Germ Line. There are three basic methodologies available for making transgenic mice, all of which involve intervention at the preimplementation stages of development (Figure 4). These techniques are: (1) the infection of preimplantation embryos with recombinant retroviruses, (2) the manipulation of embryonal stem cells (ES cells), (3) the microinjection of fertilized one-cell eggs.

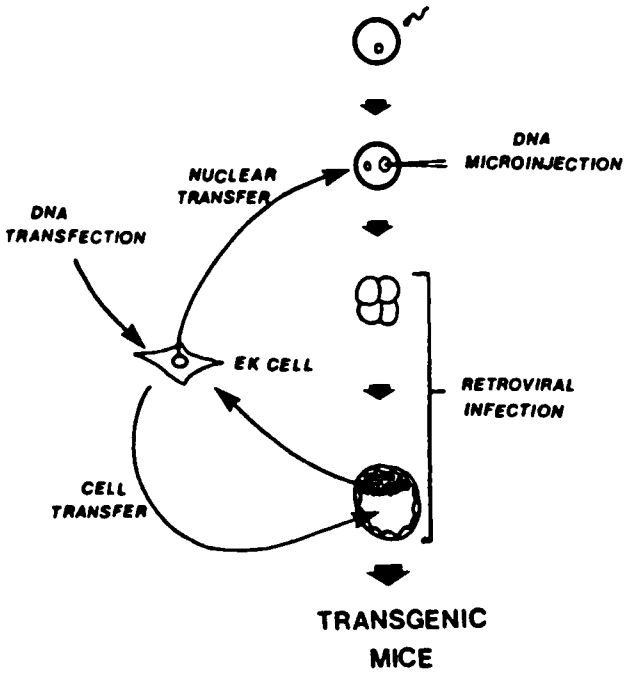


Figure 4. Routes into the germ line.

Microinjection is a quick and reliable technique that has already proved its worth, whereas retrovirus and ES cell technologies are still at the development stage with several problems still be overcome.

Retroviruses. The infection of preimplantation embryos with recombinant retroviruses is an easy technique that does not require any expensive equipment. However, each gene of interest must first be recloned into a viral vector which is then transfected into a cell line that elicits infectious recombinant virus particles. There is a limit to the size of insert that a recombinant virus can accommodate (about 10 kb) and the DNA of interest may contain splicing or termination signals that could interfere with viral function and viability. Eight-cell-stage embryos are stripped of their zona pellucidi and placed in tissue culture dishes containing fibroblasts that are producing the recombinant virus. Following infection, the embryos which have developed as far as the blastocyst stage are returned to the uterus of a pseudopregnant foster mother. A proportion of the embryos will continue normal development and give use to transgenic pups. Retroviral integration occurs via a precise mechanism that results in the insertion of a single, intact copy of the provirus containing the gene of interest, flanked by retroviral long terminal repeats (LTR's). This precision, resulting in single-copy inserts is the distinct advantage of the method over its rivals. However, while each integration event is a single copy, multiple events can occur in a single cell and multiple insertions can be found in each cell of the embryo. This results in a founder animal which is mosaic for a large number of insertions. Extensive outbreeding is required in order to establish pure lines suitable for the study of gene expression. Some transgenes have been expressed in mice containing recombinant retroviruses. However, the effect of the strong viral regulatory elements contained within the flanking LTR's upon the specificity of the expression remains to be determined.

ES Cells. These cells are pluripotent embryonic stem cells that are actually derived from blastocysts. ES cells can be cultured and manipulated *in vitro* then returned to the living organism by injection into the blastocyst. The ES cells will colonize the embryo and participate in normal development, contributing to all somatic tissues and sometimes to the germ line. ES cells can be transfected with cloned genes or infected with recombinant retroviruses prior to being injected into the blastocyst. The practical advantage of this technique is that it may be possible to select or screen transformed ES cells for some desirable property; for example, transgene copy number, chromosomal location, or expression. All founder animals are mosaics of cells with and without the transgenes and must be bred to produce pure transgenic lines. However, ES cells often fail to colonize the germ line, presumably because of the accumulation of chromosomal abnormalities during *in vitro* cultivation and selection. The problem of mosaicism in founder animals could be overcome by using transformed and selected ES cells as donors of nuclei. The pronuclei of one-cell fertilized eggs can be removed and replaced in transplantation experiments by donor nuclei, usually derived from other eggs. ES cell nuclei carrying a transgene may be able to colonize an enucleated egg which may then proceed with normal development. The resulting founder animal would carry the transgene in all its cells.

Microinjection. This remains the most popular and successful method for the production of transgenic mice, despite the fact that it demands precise technical skill and expensive equipment. The speed and reliability of the technique far outweigh any of its disadvantages. Within the overall process of making transgenic mice by microinjection, the physical introduction of DNA into the pronucleus of a one-cell egg is central. However, the more peripheral technologies, involving molecular methods, animal husbandry, and manipulation are no less important to the overall success of the techniques.

In the microinjection method, female mice are hormonally superovulated and mated with stud male mice about 12 hours

post coitum (pc). The oviducts of the donor females are removed and placed in culture. The fertilized one-cell eggs can often be clearly seen within a swollen part of the oviduct, the ampulla. Using fine forceps, the ampulla is torn apart and the eggs, surrounded and held together by cumulus cells, pour out in a cloud. Each cumulus mass is then digested with hyaluronidase, which separates the eggs from the cumulus cells. The eggs are then ready for injection, which can take place any time during the following 12 hours, just as long as the pronuclei are visible. Microinjection takes place under 400- to 600-fold magnification using Nomarski differential interference contrast microscopy. Nomarski optics are expensive, but allow the pronuclei of the eggs to be clearly seen. A cheaper alternative is Hoffman modulation contrast optics or even bright field optics. Eggs are held in place by gentle suction onto a holding pipette while a 1-micron injecting needle loaded with DNA solution is inserted through the zona pollucida, the egg membrane, and the nuclear membrane into one of the pronuclei. Usually, the larger male pronucleus is targeted. The DNA solution is then discharged by positive pressure and successful injection is indicated by the swelling of the pronucleus to twice its normal volume. Eggs that survive injection are then returned to the natural environment of a 12-hour pc pseudopregnant mother via an oviduct transfer operation. A proportion of the transplanted eggs will develop to term, and the transgenic mice in the litter are identified by hybridization analysis of high molecular weight genomic DNA isolated from mouse tails.

The mechanism by which injected DNA integrates into host chromosomes is unknown, but some ideas about the nature of the process have been inferred from a study of the state and organization of the inserts found in transgenic mice. About 70 percent of transgenic mice carry exogenous DNA in all of their somatic and germ cells, indicating that integration usually occurs prior to the first round of DNA replication. The remaining 30 percent of transgenic mice show some degree of mosaicism and in these animals integration must have occurred at some stage after the first round of DNA replication. The copy number of the transgene can vary considerably from one to several thousand. However, within a particular founder animal there is usually only one insert site. Multiple copies of a transgene are usually arranged in a head-to-tail tandom array within a single locus. The insert site within the host genome is probably determined randomly. Integration events have been observed in many different autosomes, on the x chromosome, and on the y chromosome.

The efficiency of producing transgenic mice by microinjection varies considerably between experiments. However, under optimum conditions, 60 to 80 percent of eggs survive injection. Of these, 10 to 30 percent implant in the pseudopregnant foster mother, proceed through normal development, and are born. Ten to 30 percent of pups are transgenic and roughly half of all transgenic animals express the foreign DNA. Both the condition of the DNA used to inject the eggs and the choice of mice used can have a profound effect on the overall efficiency of the technique and both can be carefully controlled.

It has been estimated by Murphy and other researchers that around one to two picolitres of DNA solution are discharged into the pronucleus of fertilized eggs. The most efficient concentration of DNA has been found to be between 1 and 5 µg/mL. This means that around 500 copies of a DNA molecule are being introduced into each egg. There is no correlation between the concentration of the microinjected DNA and the copy number of the resulting transgene. The DNA is injected in a buffer of 10-mM Tris-HCl, pH 7.4, 0.2-mM EDTA. Higher concentrations of EDTA and low concentrations of $MgCl_2$ are toxic to eggs. The DNA must be free of all contaminants that may possibly be toxic and particulate matter that could block the injection pipette. The size of the DNA has no effect on the efficiency of the process. The efficiency of the integration of linear DNA

fragments prepared by restriction enzyme digestion is fivefold greater than that of supercoiled molecules. Prokaryotic DNA sequences will contribute to a transgenic mouse as efficiently as sequences derived from eukaryotic mammalian sequences. However, contiguous vector-derived prokaryotic sequences can abolish or inhibit the expression of some eukaryotic transgenes such as globin and actin. In some circumstances, bacterial coding sequences (for example, chloramphenicol acetyl transferase [CAT]) are incorporated into hybrid genes and used to report on expression directed by eukaryotic promotor elements. Unlike some vector sequences, CAT sequences do not seem to inhibit the use of contiguous eukaryotic regulatory elements. CAT has been used to report on the expression of both the α A-crystallin promoter and the Rous Sarcoma LTR. Any inhibitory effect of prokaryote-derived DNA may therefore be specific to sequences contained within the commonly used lambda and pBR 322 derived vectors. As a general rule, therefore, investigators now remove all vector sequences prior to injecting cloned eukaryotic genes in order to maximize the quality, quantity, and reproducibility of transgene expression.

Choice of Mice. All animals used in transgenic mouse experiments should be healthy and fit in order to maximize the efficiency of the process. The choice of the strain of mouse is also important. The overall efficiency of transgenic mouse production following manipulations of C57 Bl/6 inbred mouse eggs and C57 Bl/6×CBA/5 hybrid eggs has been compared. A number of parameters were shown to be strain dependent, such as the yield of eggs from the donor female and the survival of eggs following the insult of injection. Overall, the experiments on the hybrid eggs were eightfold more efficient that those on the inbred eggs. Murphy microinjects the F2 zygotes resulting from matings between CBA/J×C57Bl/10 Fl hybrid females and stud males and achieves excellent efficiencies. Inbred zygotes should only be used when the genetic background of the host animal needs to be carefully controlled--for example, when studying immunological phenomenon such as the major histocompatability complex.

Transgene Expression. A large amount of data on the expression of exogenous DNA in transgenic mice has accumulated over the past few years. It has been found that about half of all transgenic mice made do not express their transgenes. This is thought to be either due to the presence of inhibitory suquences within the transgene (for example, prokaryotic sequences--see above) or the site of integration of the transgene. The exogenous DNA may be integrated into a chromosomal location that is transcriptionally inactive. Most transgenic mice that express the transgene do so in a manner that is appropriate to the regulatory elements present. However, the adjacent cellular DNA can influence the expression of a transgene. Some transgenic mice express their foreign DNA in an inappropriate manner, and this is thought to be due to insertion effects such as the juxtaposition of the transgene and endogenous enhancer elements.

Application of Transgenic Technology. The benefits of transgenic technology will initially be in areas of basic science. Studies on transgenic mice will contribute to most areas of mammalian biology in the following, not mutually exclusive ways:

- Gene expression studies. The analysis of the *cis*-acting regulatory sequences that mediate the tissue-specific, developmental, and physiological regulation of gene expression.
- Physiological studies. The analysis of the physiological consequences to the whole organism of the inappropriate or altered expression of normal or mutated gene products, including oncogenes.
- Genetic studies. The isolation and characterization of novel mutants. Recessive mutants often arise in transgenic mice as a consequence of the insertion of a transgene into a functional gene. The mutated gene can then be readily cloned using the transgene tag as a probe. In the future, it may

be possible to direct mutations to specific genes, either by introducing antisense constructs or by targeting of *in vitro* mutagenized constructs into ES cells using homologous recombination.

According to Murphy, out of such fundamental studies will emerge the potential for a more practical application of transgenic technology to biotechnology and agriculture. Already, transgenic pigs, sheep, and rabbits have been constructed, but it will be a long time before the agriculturalist can seriously consider transgenic technology as a new tool in animal breeding.

5 EXPRESSION AND SECRETION OF FOREIGN POLYPEPTIDES IN YEAST

Yeast is a useful organism for the expression of some eukaryotic proteins that are difficult to produce and recover in a suitable biological form from *E. coli*. The expression and secretion of eukaryotic genes in yeast as well as the advantages and limitations of the use of yeast were discussed by B. Carter (Searle Research and Development, G.D. Searle and Co., Inc., Chicago, Illinois).

The yeast organism *Saccharomyces cerevisiae* is the second-best characterized orgamism next to *E. coli*. For the purposes of heterologous gene expression, *S. cerevisiae* offers the physical and genetic ease of manipulation characteristic of microbes. It also exhibits much of the membrane-based organelle structures and protein processing systems which characterize the higher eukaryotic cell types. Particularly important in this respect is the presence of an exocrine secretion system in yeast as most of the "therapeutic" proteins which are candidates for microbial expression are secreted from their mammalian cell of origin. During their formation such proteins are subjected to a variety of intracellular environments to which cytoplasmic proteins are not exposed. Thus, secreted proteins undergo cotranslational transport across the endoplasmic reticulin, proteolytic processing, glycosylation, and a host of post-translational modifications. It is highly likely that such proteins have an amino acid composition which both adapt them to, and protects them from, the unique conditions of their formation. Many proteins require exposure to the secretory pathway if they are to be produced from yeast in an active form (for example, prochymosin, tissue plasminogen activator, etc.). Secreted proteins have an unusually high frequency of cysteine residues and hence disulphide-bond-generated secondary structures, which can cause problems when such polypeptides are produced in an appropriately reducing environment. Yeast, therefore, offers an alternative environment for heterologous protein production. The introduction of cloning systems for *S. cerevisiae* in the late 1970's was followed by a massive attempt to exploit yeast as a host for the production of mammalian proteins. In practice, the 2-micron-based shuttle vectors have been the most popular choice as vehicles for foreign gene expression in yeast. These plasmids exhibit high transformation efficiency, relatively high stability, and high copy number. Although these features make the 2-micron-based vectors relatively simple to manipulate, it should be noted that high copy number is only desirable if it is unaccompanied by biological consequences such as markedly reduced growth rate of transformant or saturation of control elements for promoters.

In attempting to maximize the yield of a particular foreign protein from yeast, it is important to maximize mRNA levels for that protein and its signal sequences, if any. The obvious route to achieve this is to drive expression of the relevant gene from a strong yeast promoter on a multicopy plasmid, and to ensure efficient transcription termination by placing a known yeast "terminator" sequence 3' to the translational terminator sequences. The promoters from genes involved in glycolysis have been strong candidates for this role as their natural products can represent as much as 5 percent of the total cell protein when the relevant gene is present as a single copy per haploid genome. For example,

when the entire phosphoglycerate kinase (PGK) gene is present on a multicopy plasmid, the PGK enzyme can constitute 80 percent of the total cell protein. The same promoter on similar vectors, however, produces chymosin at 5 percent and interferon-α (IFN-α) at 2 percent of the total cell protein. A yield of 25×10^6 I.U./L of mature IFN-α driven by the PGK promoter has been reported. This represents approximately 1 milligram of active material per litre of cells, or less than 0.05 percent of the total cell protein. Thus, the performance of the PGK promoter depends on the particular protein which is being produced.

Carter and his group have found that the pyrurate kinase (PYK) promoter gives a similar yield of interferon-α to that reported for the PGK promoter. The yield of interferon-β (IFN-β) from this promoter on multicopy vectors, however, is only 1 percent of that obtained for IFN-α_2. Furthermore, a substitution of amino acids 36-48 of IFN-β by amino acids 34-46 of IFN-α, which creates a "hybrid" interferon, is expressed at levels comparable to IFN-α_2 in constructs which vary by only 12 amino acids from natural IFN-β. This is a particulary dramatic example of what is clearly a general phenomenon--that is, that final protein levels obtained from a construct are heavily dependent on the protein itself. The mechanisms which underlie the apparent underperformance of "strong" yeast promoters when driving heterologous genes may be rather subtle, as indicated by the above example of the "hybrid" interferon, where there is no obvious structural change in the molecule to explain the discrepancy between its behavior and that of natural IFN-β.

Clearly, there are many parameters which may vary to contribute to lower than expected productivity from high copy number and strong constitutive promoter constructs. These include transinfection and translation efficiency of a particular construct, mRNA and protein stability, and vector copy number itself. A study was carried out comparing these factors in constructs where the PGK promoter was used to express either PGK itself or IFN-α. It was concluded that although both message and protein turnover were affected by the presence of the heterologous gene, the major reason for low expression of IFN-α was a reduction in mRNA synthesis. This led to the suggestion that the PGK structural gene contained a "downstream activity sequence" (DAS). However, Carter stated that this idea is not compatible with the observation that different heterologous proteins show different levels of expression. The removal of an homologous DAS should affect the expression of all foreign genes equally, and it clearly does not.

Carter then discussed the question of what does limit heterologous gene expression. He stated that, firstly, it is obvious that high-level constitutive expression of any gene on a multicopy plasmid must represent a significant metabolic strain to a transformed cell. With both yeast-derived and heterologous material this strain may be accompanied by toxic effects which can vary in both mechanism and intensity in a protein-dependent manner. Slow transformation and poor growth rate of "expressing" clones are such ubiquitous and expected phenomena that they usually receive only tangential references in the literature. Nonetheless, the slow growth of "expressing" clones obviously lends a selective advantage to cells in a population which have decreased expression of the target genes, by whatever mechanism.

Carter and his group have observed that the insertion of the entire natural PYK promoter, gene, and terminator unit into the 2-micron-based plasmid pJDB207 dramatically reduces the stability of transformants when grown under conditions which do not select for plasmid maintenance. If transformants are grown selectively to maintain LEU2 copy number (which is particulary high in pJDB207 derived plasmids), Carter found that rearrangements occur which eliminate the PYK expression capability of a proportion of the plasmids in the population. Both of these phenomena are expected if overexpression of the PYK gene itself has a deleterious effect on growth state. When the PYK promoter is used to drive the

expression of heterologous genes, the effect on growth rate is very pronounced in some cases, and these are always the instances in which yield of the target protein is either low or erratic. Lyk-promoter-IFN-α_2 plasmids generate transformants on sorbitol-stabilized selective medium with 4 days of incubation. Transformants typically yield 1 to 2 percent of total cell protein as IFN-α_2 under optional conditions. PYK-IFN-β constructs, in contrast, require 10 to 14 days to produce transformed colonies and at best produce 0.05 percent of total cell protein as IFN-β. Initial streaks of PYK-IFN-β transformants on selective medium show sparse growth and are overgrown by papillae of strong growing cells which invariably show markedly reduced IFN-β production. In this case, IFN-β is clearly more deleterious to transformant growth rate than IFN-α_2. The "hybrid" interferon described above resembles IFN-α_2 in the growth characteristics of transformants as well as in yield. The use of conditional promoters to drive heterologous gene expression should help to overcome some of the problems associated with the glycolytic promoter systems.

6 ENGINEERING SPECIFIC GENES

Tissue-type plasminogen activator (t-PA) is a thrombolytic agent with potential for improving the prospects for recovery from acute myocardial infarction. Studies on the isolation and expression of the entire t-PA gene as well as the crucial contribution that recombinant DNA technology is making toward understanding and isolating this agent was discussed by M. Browne (Biosciences Research Centre, Beecham Pharmaceuticals, UK).

Acute myocardial infarction (AMI), otherwise known as a heart attack, is a major cause of mortality and morbidity. One of the main precipitating events in AMI is the formation of an occluding thrombus in a coronary artery which then prevents blood flow to the myocardium. It is now widely accepted that early lysis of such clots will restore blood flow and hence prevent or reduce myocardial damage. There are potentially several ways of promoting clot lysis, one of the most attractive is to exploit the endogenous thrombolytic system. Thus, in normal circumstances there is a balance between clot formation (coagulation) and dissolution (thrombolysis). As a result, there is a significant reservoir of many relevant proteins, or their precursors, available in the plasma. Of specific interest, in this context, are the final events in thrombolysis; here the activation of plasminogen by protease-nicking releases the active enzyme, plasmin--a serin protease--which, in turn, lyses the cut. Activation of the plasminogen may be mediated by a number of exogenous or endogenous proteins (streptokinase, urokinase, eminase [a modified plasminogene], streptokinase complex, prourokinase, and t-PA). The first two proteins are already in clinical use; the others are being clinically evaluated. One of these proteins, t-PA, is a human enzyme normally synthesized in very small amounts by cells lining the blood vessels. Administration of large amounts of t-PA to the victim of a heart attack to promote removal of the blood clot is clearly an appealing therapeutic option.

Although the existence of t-PA has been known for some years it was originally available only in minute amounts which were extracted from perfused cadavers or human uteri. The supply situation was somewhat improved after the identification of the human Bowes melanoma cell line which constitutionally secretes t-PA at a level sufficient to enable initial clinical studies. *In vitro* studies with this material were also able to show the t-PA has two very important properties: (1) it appears to bind tightly to the fibrin found in blood clots and (2) its catalytic efficiency is markedly raised in the presence of fibrin; both properties contribute to the relative potency and selectivity of this agent. Unfortunately, even the Bowes cell line was not really productive enough to supply therapeutically significant quantities of t-PA on any scale.

At the beginning of this decade, it was realized that the most sensible approach to producing t-PA was to employ recombinant DNA technology. Many groups, including that of Browne, embarked on programs to construct and express t-PA cDNA clones. An indication of the preceived importance of t-PA is that, in addition to several academic groups, at least 40 companies in Europe, the UK, US, and Japan are currently believed to be working on t-PA. The first two groups to publish on successful cDNA cloning were at Genentech, US, and at the University of Umea, Sweden. Although the cDNA clones devised by different groups do differ in some details, the overall results obtained by most groups are very similar (none affecting the protein sequence). Translation of the cDNA sequence revealed t-PA to be a serine protease of 327 amino acids in length. The protein appears, like many serine proteases, to consist of two chains (A and B) linked by a protease-susceptible peptide bond. The 13-chain carries the catalytic center but, somewhat unusual for a serine protease, t-PA appears to be active in both cleaved and uncleaved forms. The A-chain is believed to be divided into four structural domains of unknown function (the finger, growth factor, and kringle domains). These domains appear to have counterparts in many other proteins of the coagulation and fibrinolytic systems, consistent with the idea of generation of novel proteins by shuffling of modules.

Browne and his group chose to express the recombinant t-PA cDNA in *E. coli* using the pUC 8 expression vector. *In vitro* transcription/translation revealed a ^{35}S methionine-labeled protein of about 60 kilodaltons, the expected size for unglycosylated t-PA. Recovery of active t-PA from *E. coli* carrying the pTR505 plasmid proved difficult and required the use of strong denaturing conditions combined with redox buffers to promote formation of correct disulphide bonds. Despite the fact that active t-PA of the correct size could be recovered, the yields were low. Other microbial systems for expression of t-PA (for example, yeast) have been tried by a number of groups, with some reports of success, although secretion and correct glycosylation have proved problematical in some cases. Thus, expression of t-PA in mammalian cells now appears to be the method of choice in either transient or stable systems. Browne and his group have expressed t-PA in mouse cells, and analysis of the product by zymography suggests that the product is glycosylated, unlike the *E. coli*-derived material. The basic construct used in this work (or similar types of construct) can be used in the creation of stable cell lines secreting high levels of t-PA. The most popular systems are either virus-based vectors such as BPV in mouse C127 cells, or amplification of integrated plasmids using the DHFR/methotrexate system in Chinese ovary hamster cells. Precise yields from high-level expression systems are difficult to obtain but claims are made that yields of 10 to 50 mg/L are attainable using some systems.

Browne and his group had decided at a very early stage to look at the t-PA gene itself in addition to work on cDNA. Initial attempts to clone the gene using λ phage libraries resulted in the isolation of partial clones lacking the 5' end of the gene. Browne therefore turned to cosmid technology, and rapidly isolated a clone carrying 40 kb of human DNA. The structure of the cloned DNA appeared to be very similar to that of the native (uncloned) gene as demonstrated by Southern blotting, indicating that the gene had not become rearranged during cloning. Mapping studies suggested that the B-chain coding region was located on a relatively discrete 6 kb B III fragment. The A-chain was clearly more dispersed and the 5' end of the gene appeared to be at least 10 to 15 kb away from the rest of the gene. Browne and his group were then concerned that they might not have the complete gene since there was a possibility that they might have missed further 5' introns or because control sequences (for example, the promoter) might be outside the cosmid. Rather than engage in a full mapping and sequence

program, they decided to seek evidence for cloning of the intact gene by direct expression.

Mouse L929 cells were used as the host since they do not make detectable amounts of t-PA. Stable cell lines were constructed in L929 cells using the cosmid p-TR g22. This work was facilitated by the fact that the cosmid vector carries a G418 resistance marker. Approximately half of the G418-resistant lines make a plasminogen activator. In all tests, the material was very similar to t-PA derived from Bowes melanoma cells. The cosmid thus appeared to encompass the whole t-PA genetic locus; i.e., both the coding and regulatory sequences.

Browne and his group next decided to use the cosmid to provide a cell line yielding relatively high amounts of t-PA. The cosmid was introduced into the human Bowes melanoma cells and G418 lines were obtained. One of these (TRBM6) secretes 10-fold more t-PA than the present line. By all criteria, the product appears to be the same as normal t-PA. More detailed studies have shown that these cells have a 10-fold highter t-PA-specific mRNA concentration than Bowes cells and contain 6 to 10 extra copies of the t-PA gene. By this somewhat unorthodox route, Browne and his group have succeeded in producing high levels of t-PA from a human cell line.

With a fully operational gene in their possession, Browne and coworkers investigated various aspects of its structure. The presumed promoter region was subcloned into M13 bacteriophage and sequenced. The promoter had the expected TATA and CAAT boxes. Also, a further TATA box (of unknown function) well upstream of the CAAT box was identified. The fact that this structure is a real promoter was confirmed by a group at Biogen Company (Switzerland) who were able to use it in the construction of an expression vector. The t-PA gene has discrete exons encoding the finger and growth factor domains whereas the kringles are each split into two exons (each at the same relative point). Comparison of the exon arrangement for the growth factor, kringle, and B-chain (serine protease) coding regions of t-PA with u-PA supports the concept of the 'modular' evolution of these proteins by "exon" shuffling. One obvious difference between the genes is that the introns in the t-PA gene are much larger than those in u-PA, particularly those in the 5' untranslated region. As a result, although the proteins are fairly similar in size (527 amino acids for t-PA; 411 amino acids for u-PA), the gene for t-PA is 33 kb long whereas that for u-PA is only 6.4 kb. The reason for this difference is unknown at the present time.

For the future, there are groups already working on the potential for improving t-PA. In order to do this, it is necessary to have some understanding of the structure/function relationships in the t-PA molecule, and this is now being carried out.

The role of recombinant DNA technology in the study of t-PA involves the following: (1) deduction of the primary structure of t-PA; (2) supply of material for clinical evaluation; (3) isolation and analysis of gene/evolutionary relationships; and (4) study of structure-function relationships/improved agents.

7 CONCLUSION

The development of techniques for cloning segments of DNA has revolutionized biology. As a result there has been remarkable progress in our understanding of gene organization and expression, and this has furthered the continuing development of genetic engineering, which was the aim of this European conference. The presentations have shown that it is now possible to introduce cloned genes into whole animals and plants--possibilities which still have to make their full impact. It also covered the advantages and disadvantages of expressing cloned genes in bacteria, yeasts, and cultured mammalian cells as well as studies of specific genes, the products of which have significant clinical potential.

8 REFERENCES

Carle, G.F., M. Frank, and M.V. Olson, "Electrophoretic Separations of Large

DNA Molecules by Periodic Inversions of the Electric Field," *Science*, 232 (1986), 65.

Poustka, A., and H. Lehrach, "Jumping Libraries and Linking Libraries: The Next Generation of Molecular Tools in Mammalian Genetics," *Trends in Genetic Science*, 2 (1986), 174.

Schwartz, D., and C.R. Cantor, "Separation of Chromosome-Sized DNAs by Pulsed Field Gradient Gel Electrophoresis," *Cell*, 37 (1984), 67.

Silver, J., and C.E. Buckler, "Statistical Considerations for Linkage Analysis Using Recombinant Inbred Strains and Backcrosses," *Proceedings of the National Academy of Science*, 83 (1986), 1423.

Smith, C.L., P.W. Warburton, A. Goal, and C.R. Cantor, "Analysis of Genome Organization and Rearrangements by Pulsed Field Gradient Gel Electrophoresis," *Genetic Engineering*, 8 (1986), 45.

END

7-87

DTIC